Scoring Methods in Script Concordance Tests: An Exploratory Psychometric Study

Marie-France Deschênes, PhD; Marc-André Maheu-Cadotte, PhD; Guillaume Fontaine, PhD; and Éric Dionne, PhD

ABSTRACT

Background: Despite the increasingly popular role of script concordance test (SCT) scoring methods in the evaluation of clinical reasoning, studies examining these methods in nursing are relatively scarce. This study explored the psychometric properties of five SCT scoring methods. **Method:** An SCT was administered to 12 experts and 43 learners. Scores were calculated using five methods and descriptive statistics. Differences in scores were assessed with the Mann-Whitney U test, and Spearman correlation coefficients were calculated for the different methods. **Results:** The median scores of both experts and learners differed substantially according to the scoring method used. Learners' scores were statistically different from experts' scores ($p < .01$) for each method. Spearman coefficients (range, 0.44 to 0.95) were positive for the different methods. **Conclusion:** Further research is needed to refine the influence of SCT scoring methods for use in certifying assessment of clinical reasoning in nursing. **[*J Nurs Educ*. 2023;62(10):549-555.]**

In nursing education programs, significant efforts are made to foster the development of clinical reasoning. Clinical reasoning is a competency representing the range of cognitive and metacognitive processes necessary for making sound clinical decisions (Gonzalez et al., 2021; Jessee, 2018; Simmons, 2010). The development of clinical reasoning is essential to prevent adverse events that can negatively affect the safety and quality of care (Gonzalez et al., 2021; Jessee, 2018; Richmond et al., 2020). Therefore, to ensure safe practice, developing and assessing clinical reasoning is essential.

To robustly assess all components of clinical reasoning in learners, educators are encouraged to use different but complementary tools (e.g., multiple-choice questions, objective structured clinical examinations, and oral examinations) (Brown Tyo & McCurry, 2019; Daniel et al., 2019). Among the assessment tools, script concordance tests (SCTs) are used in many health education programs, including nursing programs (Deschênes et al., 2021; Sommers, 2018). SCTs are used to assess a specific component of clinical reasoning—the ability to interpret clinical information under conditions of uncertainty (Daniel et al., 2019).

BACKGROUND

SCTs are based on the Theory of Scripts, which refers to networks of structured and organized knowledge in nurses' long-term memory (Vreugdenhil et al., 2022). Nurses with well-developed scripts can effectively mobilize knowledge to understand a situation, identify why it occurred, state hypotheses to resolve it, and initiate appropriate clinical actions in response (Vreugdenhil et al., 2022). Based on the hypotheses generated, nurses' scripts are activated and are oriented in data collection, which can reinforce, minimize or prioritize the hypotheses. SCTs are tools constructed to mimic this hypothetical deductive process of nurses' clinical reasoning (Deschênes et al., 2021).

SCTs typically include 20 to 30 clinical vignettes, each consisting of a short situation followed by independent items. The situation is deliberately ill-defined and contains elements of

Marie-France Deschênes, PhD, is an Assistant Professor, University of Montréal. Marc-André Maheu-Cadotte, PhD, is the Executive Director, Québec Network on Nursing Intervention Research. Guillaume Fontaine, PhD, is an Assistant Professor, McGill University. Éric Dionne, PhD, is a Professor, University of Ottawa.

Address correspondence to Marie-France Deschênes, PhD, Faculty of Nursing, Université de Montréal, C.P. 6128, succ. Centre-Ville, Montréal, Québec, Canada, H3C 3J7; email: marie-france.deschenes@umontreal.ca.

Disclosure: Marie-France Deschênes discloses a postdoctoral fellowship scholarship from the Social Sciences and Humanities Research Council of Canada (SSHRC #756-2021-0578). The remaining authors have disclosed no potential conflicts of interest, financial or otherwise.

Received: July 12, 2022; Accepted: February 1, 2023

doi:10.3928/01484834-20230815-05

ambiguity, uncertainty, or incompleteness (Dory et al., 2012; Lubarsky et al., 2013). As shown in **Figure 1**, a typical item contains three parts: "If you think…and then you notice…your hypothesis becomes…" Part 1 provides a hypothesis. Part 2 presents a new element of information, such as findings of a physical or mental assessment, a test result, a preexisting condition, or verbalizations expressed by the patient. Part 3 provide a scale describing the effects of the new information on the suggested hypothesis.

Evidence-based recommendations exist to guide the construction of SCTs, but there is ongoing debate about scoring methods (Daniel et al., 2019). Psychometric qualities published on SCTs are based mainly on the measurement of the internal consistency of the tests, expressed by Cronbach's alpha coefficient, which is found to be high using a substantial number of items (60 to 90 items). The differences in scores between experts and learners supported the tools' construct validity (Dory et al., 2012; Lubarsky et al., 2013).

The methods for establishing scores in SCTs are complex and varied but usually involve two steps (Dionne et al., 2017). First, a group of experts individually answer SCT items without consulting peers or references. Second, the frequency distribution of the experts' response categories to SCT items are used to determine the learners' scores, hence the notion of "concordance." The score calculated for each item reflects the degree of agreement between the learners' response and those of the experts (Dory et al., 2012; Lubarsky et al., 2013), even though not all experts agree on the choice of response categories. Because of the variability in expert responses, the SCT does not typically involve "right" or "wrong" responses to items (leading to dichotomous scores). However, patterns sometimes may emerge in experts' responses.

The aggregate scoring method is used for determining SCT scores and has been recommended repeatedly (Dory et al., 2012; Lubarsky et al., 2013). In this method, which was proposed for SCTs by Charlin et al. (1998), learners get the maximum score (1) if they choose the modal category determined by experts, whereas they get partial credit if they make the same answer choice as at least one expert. If they make an answer choice that no expert has chosen, they get no points (0). **Table 1** illustrates the steps in determining scores with this method, using the example of a fictitious panel comprising 10 experts. However, the aggregate scoring method has been criticized by some authors (Bland et al., 2005; Lineberry et al., 2013; Wilson et al., 2014). Bland et al. (2005) questioned the robustness of the SCT for assessing clinical reasoning, given the absence of a single correct answer for items. They argued that this scoring method does not consider the accuracy of the learners' choice of response categories. For example, an item in which all experts answered "strengthened" is assigned the same score of 0 for learners who answered "very strengthened" and "very weakened."

Some studies have tested alternative SCT scoring methods to explore the psychometric qualities of the SCT (Bland et al., 2005; Exantus, 2020; Wilson et al., 2014). Different scoring methods of SCTs include: (1) the dichotomous score method with five response categories; (2) the dichotomous score method with three response categories; (3) the aggregate scoring method; (4) the method that penalizes distance from the experts' modal choices (DFEMC); and (5) the combined aggregate scoring and DFEMC method. The psychometric properties of these methods have been evaluated in various studies in the medical field (Bland et al., 2005; Exantus, 2020; Wilson et al., 2014). Based mainly on internal consistency measures obtained in these samples, researchers have recommended the use of different scoring methods to evaluate clinical reasoning. Thus, based on the findings, it is still difficult to suggest one scoring method over another. Nursing studies on the use of SCTs for clinical reasoning assessment have, with one exception, always used the aggregate scoring method to determine learners' scores (Blanié et al., 2020; Dawson et al., 2014; Deschênes et al., 2011). Only Dionne et al. (2017) have discussed using dichotomous scores in a secondary analysis of SCT data in nursing (Latreille, 2012).

Studies to examine alternative scoring methods and their psychometric properties remain scarce in nursing education. As SCTs increasingly are being introduced into educational programs to teach and assess clinical reasoning, empirical evidence is needed to support these educational choices. Thus, this exploratory study examined the psychometric properties of five SCT scoring methods (**Figure A**; available in the online version of the article) based on expert and learner scores.

METHOD

Study Design

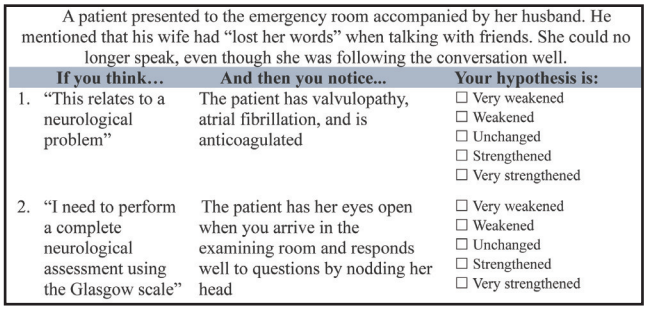
This exploratory psychometric study examined different methods of scoring SCTs and their psychometric properties. After completing the SCT items, expert and learner scores were calculated using five methods: (1) the dichotomous score method with five response categories; (2) the dichotomous score method with three response categories; (3) the aggregate scoring method; (4) the DFEMC method; and (5) the combined aggregate scoring and DFEMC method.

Context and Participants

The study was conducted in an educational institution in Canada. The university provides undergraduate nursing education to more than 200 students per year. This 3-year program is based on a competency-based approach and consists of 103 credits; one credit is equivalent to 45 hours of educational activities. Participants in the study were experts and undergraduate nursing learners. To be considered experts, those recruited had practiced for at least 5 years as a nurse in a general practice setting. These experts also had to hold a nurse educator role and had to be designated by their peers as someone known to exercise sound clinical reasoning. They were recruited using purposive sampling from local resources (i.e., professors with academic and research functions, lecturers, and laboratory coordinators) at the university. Undergraduate nursing learners were recruited through convenience sampling. The only inclusion criterion for learners was current enrollment in the university's undergraduate nursing program.

Research Activities

Participant sociodemographic data were collected using an online survey on the web-based platform used for the SCT. The experts answered the SCT items in a Microsoft® Word docu-



A patient presented to the emergency room accompanied by her husband. He mentioned that his wife had "lost her words" when talking with friends. She could no longer speak, even though she was following the conversation well.

If you think…	And then you notice...	Your hypothesis is:
1. "This relates to a neurological problem"	The patient has valvulopathy, atrial fibrillation, and is anticoagulated	☐ Very weakened ☐ Weakened ☐ Unchanged ☐ Strengthened ☐ Very strengthened
2. "I need to perform a complete neurological assessment using the Glasgow scale"	The patient has her eyes open when you arrive in the examining room and responds well to questions by nodding her head	☐ Very weakened ☐ Weakened ☐ Unchanged ☐ Strengthened ☐ Very strengthened

Figure 1. Example of a vignette from a script concordance test.

ment, as the web platform was not accessible for optimal use of the SCT question formats. The learners answered the items on a web platform (Theia) in an asynchronous mode during a 7-week period outside of class or work hours and in several sessions according to their availability. The web platform could be used on a computer, tablet, or smartphone. The primary study researcher was available by email or phone to answer questions.

The Script Concordance Test

The SCT used in this study contained 81 items related to general medical and surgical nursing, including 53 items on clinical aspects (e.g., interpretation of laboratory results or physical assessment and decisions regarding the administration of medication) and 28 items on the therapeutic relationship (e.g., interpretation of patients' verbalizations and caring interventions such as showing empathy and respect). The steps for developing an SCT, as recommended by Lubarsky et al. (2013), were observed. During the validation process of the SCT, a Delphi method was used (Keeney et al., 2011) to determine, among the primary researcher's vignettes initially written, the vignettes that were deemed the most relevant for learners. The SCT development and validation steps, which have been published elsewhere (Deschênes & Goudreau, 2020), preceded the present study.

This study used the SCT for learning purposes (assessment for learning). When learners provided answers to SCT items on the web platform, they benefited from automated feedback that presented the experts' reasoning that led to their decisions (i.e., experts' answers for each item and a brief justification that explained their choices). Learners could validate their level of concordance with the experts' answers and identify knowledge gaps. Missing data were considered as no response to the item. Considering that missing data can influence the inferences made surrounding the performance of the participants and the psychometric analysis of the items, SCT therapeutic relationship items were selected for data analysis in this study. The percentage of learners' response rates was higher in the therapeutic relationship component; 38 learners (88%) answered all of the items in this component compared with 31 learners (73%) in the clinical activities component. For the experts' response rates, there were no missing data in either component.

Data Collection

Data collected related to expert and learner response category choices for therapeutic relationship items (n = 28) on the SCT. Data were downloaded from the online platform for transcription into Excel® spreadsheets. To facilitate data processing using the analysis software, response categories were recoded as numerical values using a 5-point Likert scale ranging from 1 = *very weakened* to 5 = *very strengthened*. For the three-response dichotomous scoring method, response categories were recoded such that the very weakened and weakened responses, as well as the very strengthened and strengthened categories, were combined, resulting in the following three response categories: negative, neutral, or positive. From the compilation of the frequency of the experts' response categories for each item, the score for each response category was calculated according to the different scoring methods.

TABLE 1
Steps for Determining Scores in a Script Concordance Test Using the Aggregate Scoring Method

	Response Categories				
Step	**1**	**2**	**3**	**4**	**5**
1. Identify the number of experts in each response category to determine the modal category	5	3	1	1	0
2. Divide the number of experts in each response category according to the number of experts who endorsed the modal response (here 5)	5/5	3/5	1/5	1/5	0/5
3. Determine the score for learners in each response category	1	0.6	0.2	0.2	0

Note. 1 = very weakened; 2 = weakened; 3 = unchanged; 4 = strengthened; 5 = very strengthened.

Determination of Scores and Analysis

Based on the 12 experts' choices of item response categories, the five SCT scoring modalities were compared in the Excel spreadsheets. The scores for each modality were used to determine the performance level of experts and learners. Expert scores were calculated excluding their own choices of item response categories to avoid biased overestimations of their scores. Learners' scores were calculated according to the experts' responses to the items for each scoring method. As with any assessment method and because participants had no time restrictions in completing the SCT, no points were awarded for missing data, even though it could have been due to an inability to answer the question, a decision to skip the question, or an inadvertent omission. This occurred for five (12%) learners but no experts.

After calculating the scores using the different methods, descriptive statistics were calculated using Excel and SPSS® version 27. Normality of score distributions was assessed using the Shapiro-Wilk test and by observing the Q-Q and box plots. As scores were not normally distributed for all setting methods,

TABLE 2
Participant Demographics (n = 38)[a]

Demographic	*n* (%)
Sex	
Male	5 (13.2)
Female	33 (86.8)
Age (years)	
≤ 20	2 (5.3)
21 to 25	32 (84.2)
26 to 30	0
31 to 40	3 (7.9)
≥ 41	1 (2.6)
Prior studies in the health and social services field	
Yes	32 (84.2)
No	6 (15.8)
Work experience in the health and social services field	
Yes	16 (42.1)
No	22 (57.9)

[a]*Five (12%) participants did not complete the online questionnaire.*

medians and interquartile ranges are reported. As an indicator of reliability, the internal consistency of the SCT items was calculated and expressed by Cronbach's alpha coefficients, which ranged from 0.00 to 1.00. In health science education settings, Downing (2004) suggested a coefficient of .90 or higher for very high-stake, .80 to .89 for moderate-stake, and .70 to .79 for low-stake assessments. As an indicator of construct validity, scores between experts and learners for each setting method were analyzed using the Mann–Whitney U test. Sociodemographic data were analyzed using descriptive statistics. Spearman correlation coefficients were calculated for the different scoring methods to assess their convergence. Results were considered statistically significant at a threshold of $p < .05$.

Ethical Considerations

The study was approved by the University Research Ethics Board (17-156-CERES-D). Written consent to participate in the study was obtained from participants before they completed the SCT.

RESULTS

Sociodemographic Data

The 12 experts comprised professors with academic and research functions (n = 4), lecturers (n = 5), a laboratory supervisor (n = 1), an undergraduate program supervisor (n = 1), and a clinical placement supervisor (n = 1). All of the participants had experience in general care areas and had frequent contact with learners. Forty-five learners agreed to participate; two subsequently dropped out, and their response category choices were removed from the data analysis. Sociodemographic data of the 38 learners who completed the sociodemographic questionnaire are summarized in **Table 2**.

Comparison of Experts and Learners Scores Across Scoring Methods

The median scores of experts and learners differed according to the scoring method (**Table 3**). The five-response dichotomous scoring method (M1) was the most penalizing method for both groups, whereas the three-response dichotomous scoring method (M2) was the most rewarding. The aggregate scoring method (M3) was more penalizing than the DFEMC method (M4) and the combined aggregate scoring and DFEMC method (M5). Differences also were observed between the measures of the interquartile range, with the DFEMC method (M4) showing the lowest interquartile range in both groups (6.0 for experts and 4.3 for learners). All of the results were statistically significant (**Table 4**). The internal consistency coefficient for each method ranged from 0.68 to 0.84.

Spearman correlation coefficients were calculated to estimate the level of correlation between the different scoring methods. Predominantly positive (between 0.44 and 0.95) and statistically significant ($p < .001$) Spearman correlations were found across the scoring methods (**Table 5**).

DISCUSSION

Assignment of learners' scores is directly dependent on the variety of expert response category choices to items in an SCT. Thus, determining the most effective scoring method remains a critical issue when using SCTs to evaluate clinical reasoning. This exploratory psychometric study compared five different SCT scoring methods: two dichotomous scoring methods (with three or five response categories), the aggregate scoring method, the DFEMC method, and the combined aggregate scoring and DFEMC method. The results showed the median scores of both experts and learners differed substantially according to the scoring method used. However, regardless of the method used, satisfactory internal consistency indices were found, learners' scores remained statistically different from experts' scores, and positive associations were identified for each method.

In this study, the five-response dichotomous and aggregate scoring methods were the two most penalizing methods. The aggregate scoring method, developed by Norman (1985), requires that weights are derived from the performance of a group of experts who take the test under the same conditions as the candidates. This method, which is based on allowing for variability in experts' choices when determining scores, can reflect the differences and trends often found in professional practice (Norcini et al., 1990). However, other methods also can incorporate this principle. One of the criticisms of the aggregate scoring method is that it may give weight to response categories considered to be less than ideal or opposite to the direction of the response category chosen by the majority of experts (Bland et al., 2005; Exantus, 2020; Lineberry et al., 2013; Wilson et al., 2014).

TABLE 3
Descriptive Analysis of Experts and Learners' Scores According to Five Script Concordance Test Scoring Methods

Experts and Learners Scores According to Scoring Methods	M1	M2	M3	M4	M5
Experts' scores (n = 12)					
Median (interquartile range)	71.1 ± 16.6	89.3 ± 9.8	71.8 ± 8.3	88.4 ± 6.0	81.2 ± 7.7
Maximum value	82.1	92.9	91.4	92.9	90.7
Minimum value	48.9	78.6	62.4	79.6	73.4
Learners' scores (n = 43)					
Median (interquartile range)	55.0 ± 9	82.2 ± 14.3	67.4 ± 12.3	81.4 ± 4.3	74.4 ± 6.0
Maximum value	71.1	89.3	82.8	89.3	85.4
Minimum value	28.3	32.1	30.5	41.8	38.6
Alpha	0.68	0.74	0.76	0.84	0.79

Note. M1 = dichotomous scores (5 categories); M2 = dichotomous scores (3 categories); M3 = aggregate scores; M4 = distance from experts' modal choice method; M5 = combination of M3 and M4; CI = confidence interval (of median).

The five-response dichotomous and the aggregate scoring methods appear to have disadvantaged learners in this study who failed to choose the experts' modal response but made choices in the same direction as the experts. For example, they considered an intervention as "relevant" when the experts' modal response was "very relevant." Therefore, scores were higher using the DFEMC method and the three-response dichotomous scoring method (comprising positive, neutral, and negative response categories). These results are consistent with those of Bland et al. (2005), who compared the psychometric properties of five SCT scoring methods. The researchers found the reliability and validity of the dichotomous scoring methods were similar to the five-choice aggregate scoring method. They based their arguments on the internal consistency indices of the SCT scoring methods and the significant intergroup differences between experts and learners.

Our results indicate the DFEMC method is relevant for SCT scoring. As Wilson et al. (2014) proposed, this study investigated the combined aggregate scoring and DFEMC method. Although rarely explored, this combined method allows for variability in expert response category choices and awards partial credits based on the direction of learners' response category choices. Thus, this method accounts for situations in which learners make a response choice in the same direction as the majority of experts; these learners receive more partial credits than learners who choose in the opposite direction. Similar to Wilson et al. (2014), we believe these two methods resonate with the philosophy underlying the SCT, where variability can remain in response category choices even among experts. However, the low interquartile ranges of the DFEMC method (M4) raise questions about its ability to discriminate between learners' performance. In other words, if learners obtain approximately identical scores, the validity of the interpretation of the scores becomes questionable. Thus, it seems appropriate to examine SCT scores by contrasting them with other measures of clinical reasoning among the same learners.

Overall, the strong associations noted between the different scoring methods suggest that learners with a low score also will obtain a low score with another method. Nonetheless, if the SCT is to be used for certification purposes, establishing a threshold of success or an acceptable gap between learners' and experts' scores cannot be accomplished in a haphazard manner. Although some efforts have been made to investigate the creation of a pass mark (i.e., a minimum performance level) (Charlin et al., 2010; Linn et al., 2013), other methods, under different measurement theories, need to be used to refine our understanding of the quality of scoring in this type of test.

LIMITATIONS

This study was conducted in a nursing education program at a single institution with a relatively small number of learners and experts. Participation in the study was not part of the regular academic curricula, which may have limited learners' availability and interest in participating. Because this was an asynchronous online activity, it was challenging to control for potential contamination between participants or others outside the study. In a circumstance where scores would be compiled to measure learner performance within the education program, an asynchronous online test should be administered using a program to ensure the test taker is alone and there are no potential contaminants, such as open textbooks and notes on a desk.

This also was the first time such an assessment tool had been experimented in this nursing education program. Learners were unfamiliar with the SCT, which may have influenced the quality of their answers. In addition, the SCT was administered only once to participating students, limiting the ability to explore other psychometric properties of the scoring methods, such as their test-retest stability and sensitivity in detecting differences or whether clinical reasoning substantially improved. Thus, the results provide a partial picture of the validity and reliability of SCT scoring methods. However, this study contributes significantly to the literature because it offers new evidence in an area of nursing education that currently lacks firmly grounded guidelines.

IMPLICATIONS AND RECOMMENDATIONS

SCTs are tools whose originality lies, among other things, in the involvement of experts in the feedback given to learners who answer the test items. The scores established in the SCT

TABLE 4
Analysis of Scores Using Mann-Whitney and Wilcoxon Tests

Variable	Method 1	Method 2	Method 3	Method 4
Mann-Whitney U	74	116	150	53
Wilcoxon	1020	1062	1096	999
Z	–3.75	–2.91	–2.20	–4.18
p	< .001	0.004	.028	< .001

TABLE 5
Spearman Correlation Coefficients for Script Concordance Test Scoring Method

Method	Method 2	Method 3	Method 4	Method 5
1	.44*	.81*	.68*	.87*
2		.52*	.69*	.60*
3			.56*	.95*
4				.76*

*$p < .001$.

depend directly on expert responses to items, which makes the resulting evaluation of clinical reasoning contextualized. This can be a problem if the assessment is crucial for certification or a professional entrance exam. Based on our results, we believe the DFEMC method appears to be the optimal method for setting SCT scores. This method allows for variability in expert response category choices and awards partial credits based on the direction of learners' response category choices.

However, caution is recommended in assessing clinical reasoning using SCTs. Based on our results and published literature, research data are still too scarce to recommend, beyond a reasonable doubt, one scoring method over another. Further research is needed on SCT scoring methods, the stability of results over time, and the convergence of SCT scores with other measures of clinical reasoning. Given the current empirical evidence, we recommend the formative use of SCTs in nursing education to promote the learning of clinical reasoning. This allows learners to situate themselves in relation to the response's choices of experts in their field and to examine the elements when they are discordant with most experts.

CONCLUSION

In the present study, five scoring methods for script concordance tests were explored in the context of nursing education. With each method, learners' scores were statistically different from experts' ($p < .01$). Spearman coefficients (ranging between 0.44 and 0.95) were positive for the five methods. The DFEMC method was the most rewarding, whereas the five-response dichotomous scoring method emerged as the most penalizing for both groups in setting SCT scores. The three-response dichotomous scoring method reported similar results to the DFEMC method.

As with many studies on this topic, the experts were selected based on their experience and recognition by their peers. However, this does not mean that their response choices are infallible in measuring clinical reasoning. Further research is needed to clarify what the expertise of a panel in SCTs entails in developing a robust scoring system for assessing clinical reasoning. Studies investigating fidelity measurement in various forms also are required to document inter-panel, inter-expert, and test-retest measurement errors (learners and experts). Finally, further research is needed to refine our understanding of the influence of SCT scoring methods and their role in certifying the assessment of clinical reasoning.

REFERENCES

Bland, A. C., Kreiter, C. D., & Gordon, J. A. (2005). The psychometric properties of five scoring methods applied to the script concordance test. *Academic Medicine, 80*(4), 395–399. https://doi.org/10.1097/00001888-200504000-00019 PMID:15793026

Blanié, A., Amorim, M.-A., & Benhamou, D. (2020). Comparative value of a simulation by gaming and a traditional teaching method to improve clinical reasoning skills necessary to detect patient deterioration: A randomized study in nursing students. *BMC Medical Education, 20*(1), 53. https://doi.org/10.1186/s12909-020-1939-6 PMID:32075641

Brown Tyo, M., & McCurry, M. K. (2019). An integrative review of clinical reasoning teaching strategies and outcome evaluation in nursing education. *Nursing Education Perspectives, 40*(1), 11–17. https://doi.org/10.1097/01.NEP.0000000000000375 PMID:30095730

Charlin, B., Brailovsky, C., Leduc, C., & Blouin, D. (1998). The diagnosis script questionnaire: A new tool to assess a specific dimension of clinical competence. Advances in Health Sciences Education: Theory and Practice, 3(1), 51–58. https://doi.org/10.1023/A:1009741430850 PMID:12386395

Charlin, B., Gagnon, R., Lubarsky, S., Lambert, C., Meterissian, S., Chalk, C., Goudreau, J., & van der Vleuten, C. (2010). Assessment in the context of uncertainty using the script concordance test: More meaning for scores. *Teaching and Learning in Medicine, 22*(3), 180–186. https://doi.org/10.1080/10401334.2010.488197 PMID:20563937

Daniel, M., Rencic, J., Durning, S. J., Holmboe, E., Santen, S. A., Lang, V., Ratcliffe, T., Gordon, D., Heist, B., Lubarsky, S., Estrada, C. A., Ballard, T., Artino, A. R., Jr., Sergio Da Silva, A., Cleary, T., Stojan, J., & Gruppen, L. D. (2019). Clinical reasoning assessment methods: A scoping review and practical guidance. *Academic Medicine, 94*(6), 902–912. https://doi.org/10.1097/ACM.0000000000002618 PMID:30720527

Dawson, T., Comer, L., Kossick, M. A., & Neubrander, J. (2014). Can script concordance testing be used in nursing education to accurately assess clinical reasoning skills? *Journal of Nursing Education, 53*(5), 281–286. https://doi.org/10.3928/01484834-20140321-03 PMID:24641082

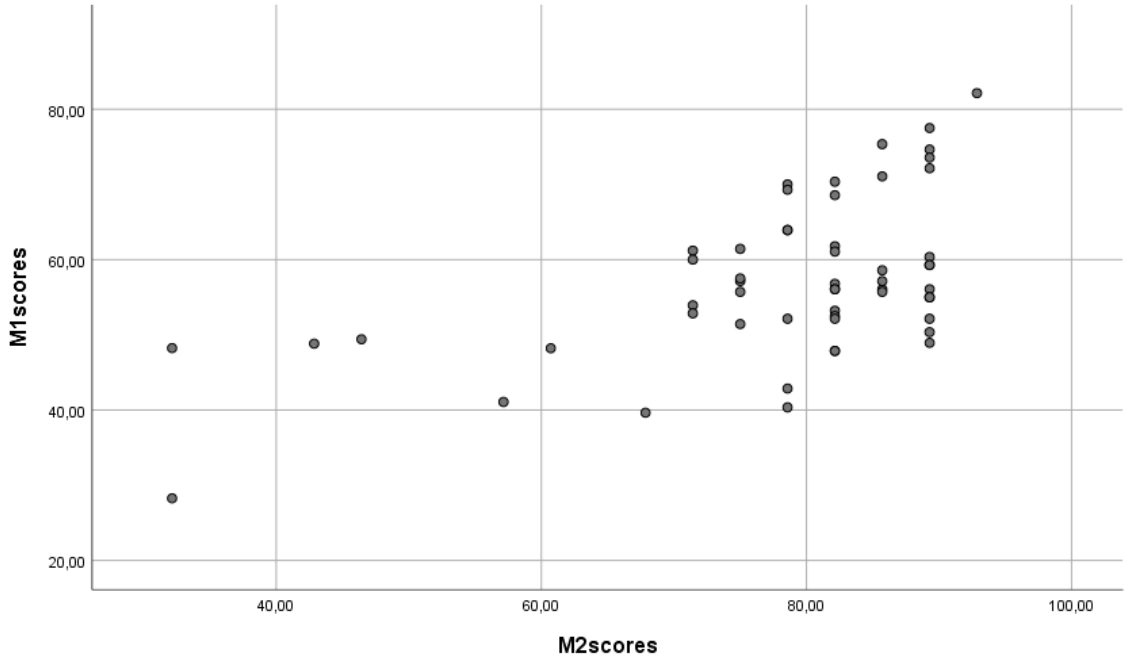
Deschênes, M.-F., Charlin, B., Gagnon, R., & Goudreau, J. (2011). Use of a script concordance test to assess development of clinical reasoning in nursing students. *Journal of Nursing Education, 50*(7), 381–387. https://doi.org/10.3928/01484834-20110331-03 PMID:21449528

Deschênes, M.-F., & Goudreau, J. (2020). L'apprentissage du raisonnement clinique infirmier dans le cadre d'un dispositif éducatif numérique basé sur la concordance de scripts [The learning of nursing clinical reasoning within the framework of a digital educational device based on the concordance of scripts]. Pédagogie Médicale [Medical Education], 21(3), 143–157. https://doi.org/10.1051/pmed/2020041

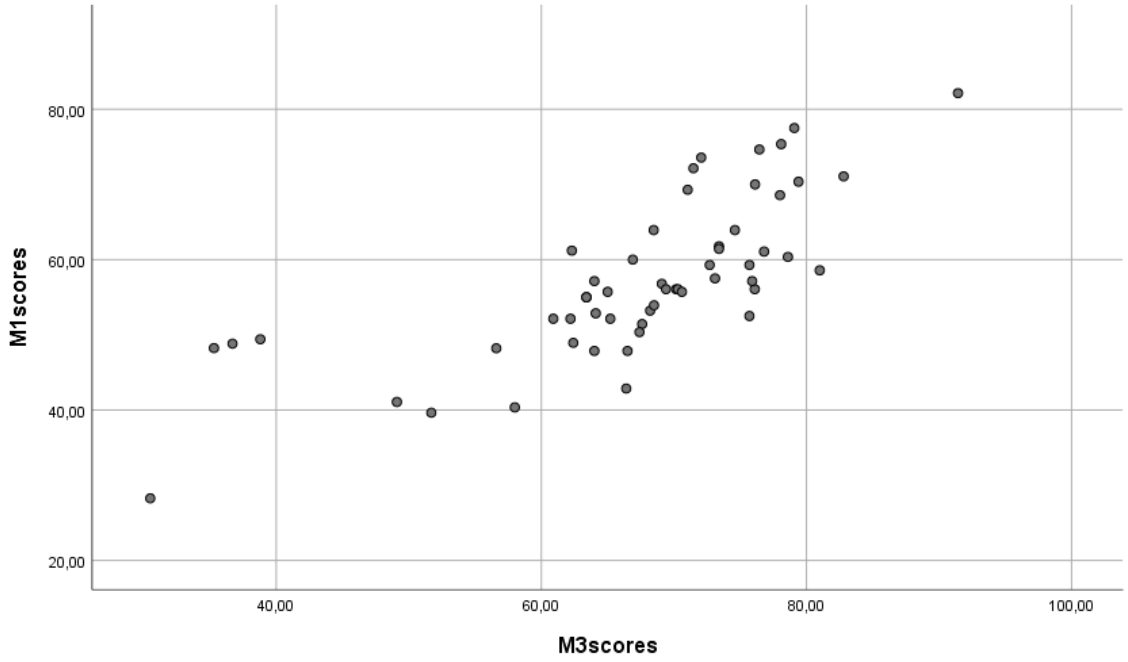
Deschênes, M.-F., Létourneau, D., & Goudreau, J. (2021). Script concordance approach in nursing education. *Nurse Educator, 46*(5), E103–E107. https://doi.org/10.1097/NNE.0000000000001028

PMID:33958554
Dionne, É., Grondin, J., & Latreille, M.-È. (2017). Exploration des scores à un test de concordance de script sous la loupe de la modélisation de Rasch [Exploring concordance test scoring from the perspective of Rasch model]. In E. Dionne & I. Raîche (Eds.), Mesure et évaluation en éducation médicale. Regards actuels et prospectifs [Measurement and evaluation in medical education. Current and prospective views] (pp. 77–110). Presses de l'Université du Québec.
Dory, V., Gagnon, R., Vanpee, D., & Charlin, B. (2012). How to construct and implement script concordance tests: Insights from a systematic review. *Medical Education, 46*(6), 552–563. https://doi.org/10.1111/j.1365-2923.2011.04211.x PMID:22626047
Downing, S. M. (2004). Reliability: On the reproducibility of assessment data. *Medical Education, 38*(9), 1006–1012. https://doi.org/10.1111/j.1365-2929.2004.01932.x PMID:15327684
Exantus, J. (2020). Comparaison des propriétés métriques des scores obtenus avec un test de concordance de script au regard de trois méthodes de détermination des scores [Comparison of psychometric properties of a script concordance test scores using three scoring methods] [Unpublished master's thesis]. University of Ottawa.
Gonzalez, L., Nielsen, A., & Lasater, K. (2021). Developing students' clinical reasoning skills: A faculty guide. *Journal of Nursing Education, 60*(9), 485–493. https://doi.org/10.3928/01484834-20210708-01 PMID:34467807
Jessee, M. A. (2018). Pursuing improvement in clinical reasoning: The integrated clinical education theory. *Journal of Nursing Education, 57*(1), 7–13. https://doi.org/10.3928/01484834-20180102-03 PMID:29381154
Keeney, S., Hasson, F., & McKenna, H. (2011). *The Delphi technique in nursing and health research*. Wiley-Blackwell. https://doi.org/10.1002/9781444392029
Latreille, M.-E. (2012). Évaluation du raisonnement clinique d'étudiantes et d'infirmières dans le domaine de la pédiatrie, à l'aide d'un test de concordance de script [Use of a script concordance test in the field of pediatrics to assess clinical reasoning of nurses and nursing students] [Unpublished master's thesis.] University of Ottawa.
Lineberry, M., Kreiter, C. D., & Bordage, G. (2013). Threats to validity in the use and interpretation of script concordance test scores. *Medical Education, 47*(12), 1175–1183. https://doi.org/10.1111/medu.12283 PMID:24206151
Linn, A. M., Tonkin, A., & Duggan, P. (2013). Standard setting of script concordance tests using an adapted Nedelsky approach. *Medical Teacher, 35*(4), 314–319. https://doi.org/10.3109/0142159X.2012.746446 PMID:23228081
Lubarsky, S., Dory, V., Duggan, P., Gagnon, R., & Charlin, B. (2013). Script concordance testing: From theory to practice: AMEE guide no. 75. *Medical Teacher, 35*(3), 184–193. https://doi.org/10.3109/0142159X.2013.760036 PMID:23360487
Norcini, J. J., Shea, J. A., & Day, S. C. (1990). The use of aggregate scoring for a recertifying examination. *Evaluation & the Health Professions, 13*(2), 241–251. https://doi.org/10.1177/016327879001300207
Norman, G. R. (1985). Objective measurement of clinical performance. *Medical Education, 19*(1), 43–47. https://doi.org/10.1111/j.1365-2923.1985.tb01137.x PMID:3969023
Richmond, A., Cooper, N., Gay, S., Atiomo, W., & Patel, R. (2020). The student is key: A realist review of educational interventions to develop analytical and non-analytical clinical reasoning ability. *Medical Education, 54*(8), 709–719. https://doi.org/10.1111/medu.14137 PMID:32083744
Simmons, B. (2010). Clinical reasoning: Concept analysis. *Journal of Advanced Nursing, 66*(5), 1151–1158. https://doi.org/10.1111/j.1365-2648.2010.05262.x PMID:20337790
Sommers, C. L. (2018). Measurement of critical thinking, clinical reasoning, and clinical judgment in culturally diverse nursing students—A literature review. *Nurse Education in Practice, 30,* 91–100. https://doi.org/10.1016/j.nepr.2018.04.002 PMID:29669305
Vreugdenhil, J., Döpp, D., Custers, E. J. F. M., Reinders, M. E., Dobber, J., & Kusukar, R. A. (2022). Illness scripts in nursing: Directed content analysis. *Journal of Advanced Nursing, 78*(1), 201–210. https://doi.org/10.1111/jan.15011 PMID:34378221
Wilson, A. B., Pike, G. R., & Humbert, A. J. (2014). Analyzing script concordance test scoring methods and items by difficulty and type. *Teaching and Learning in Medicine, 26*(2), 135–145. https://doi.org/10.1080/10401334.2014.884464 PMID:24702549

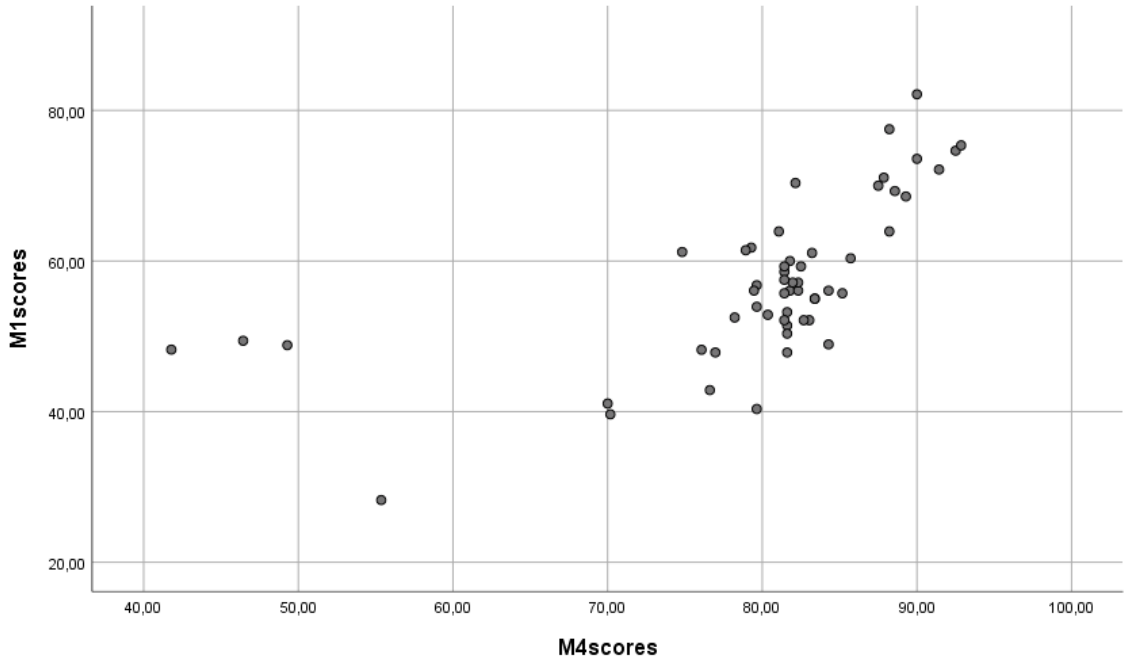
Figure A: Scatterplots of different SCT scoring methods



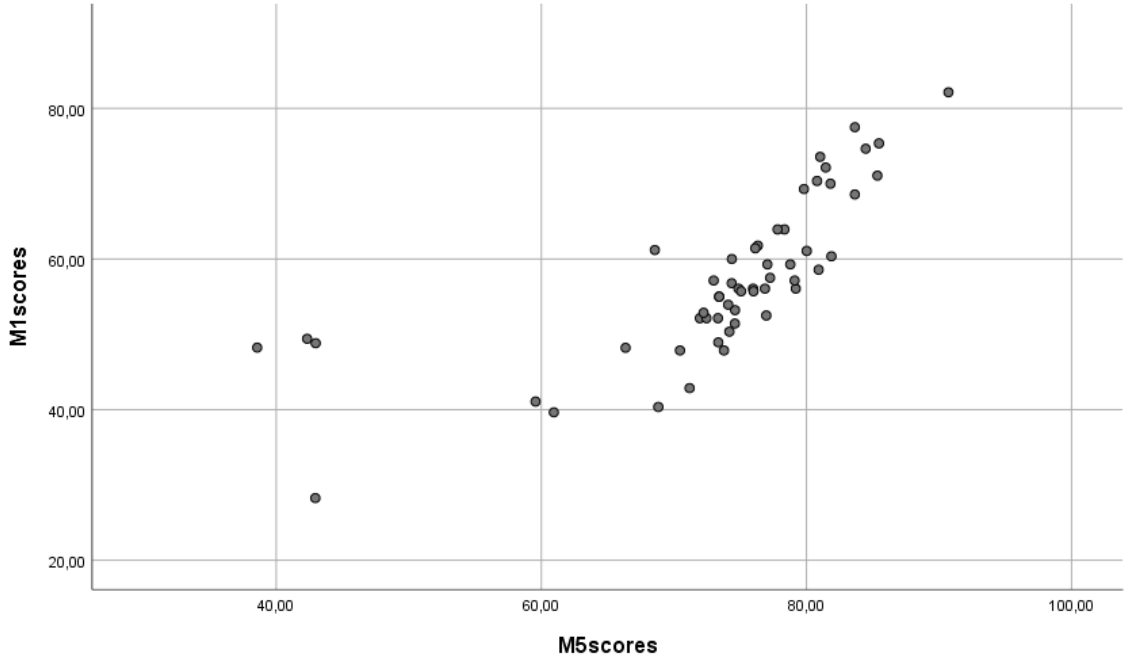
1. Scatterplot incorporating method 1 (five-response dichotomous scoring) and method 2 (three-response dichotomous scoring)



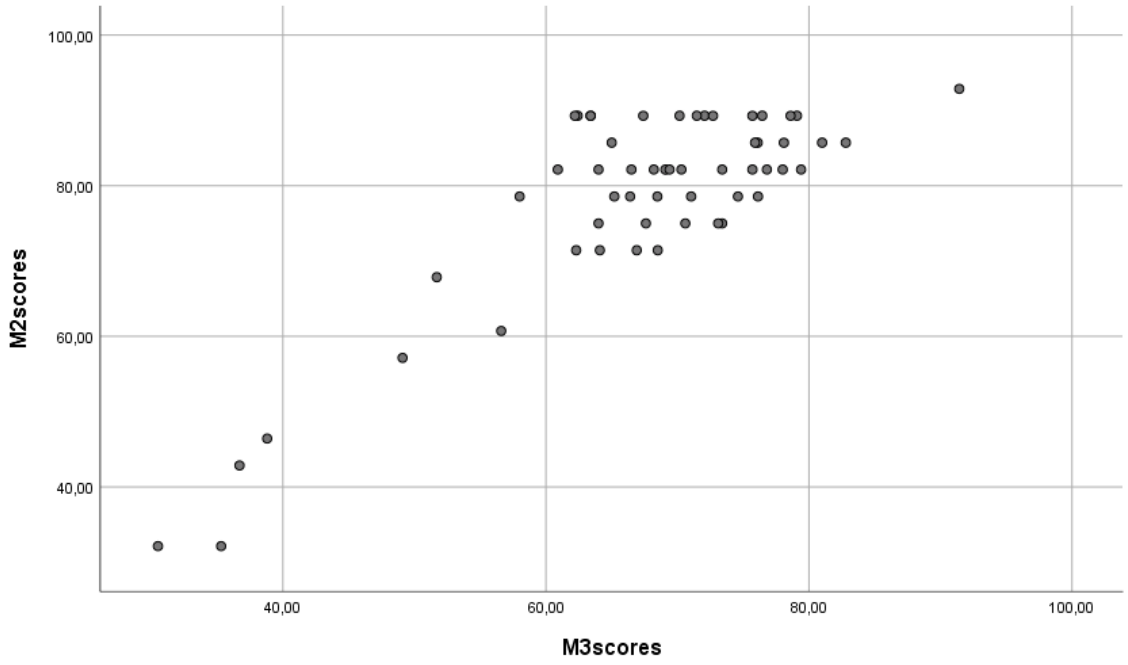
2. Scatterplot incorporating method 1 (five-response dichotomous scoring) and method 3 (aggregated scores)



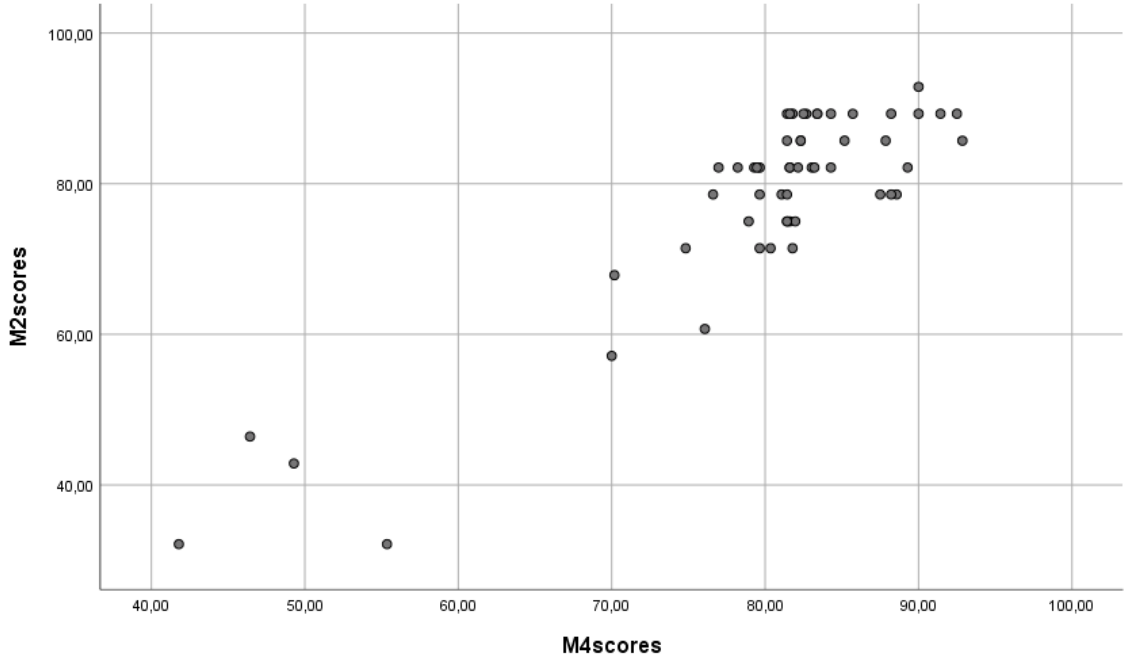
3. Scatterplot incorporating method 1 (five-response dichotomous scoring) and method 4 (distance from experts' modal choices)



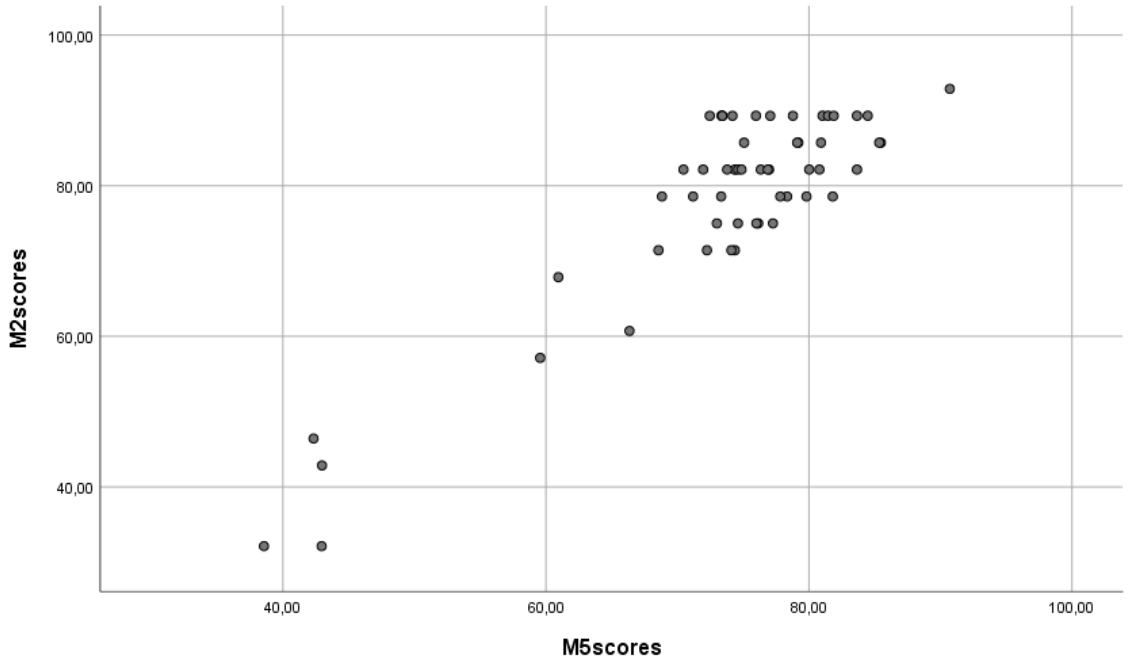
4. Scatterplot incorporating method 1 (five-response dichotomous scoring) and method 5 (combination of aggregate scores and distance from experts' modal choices)



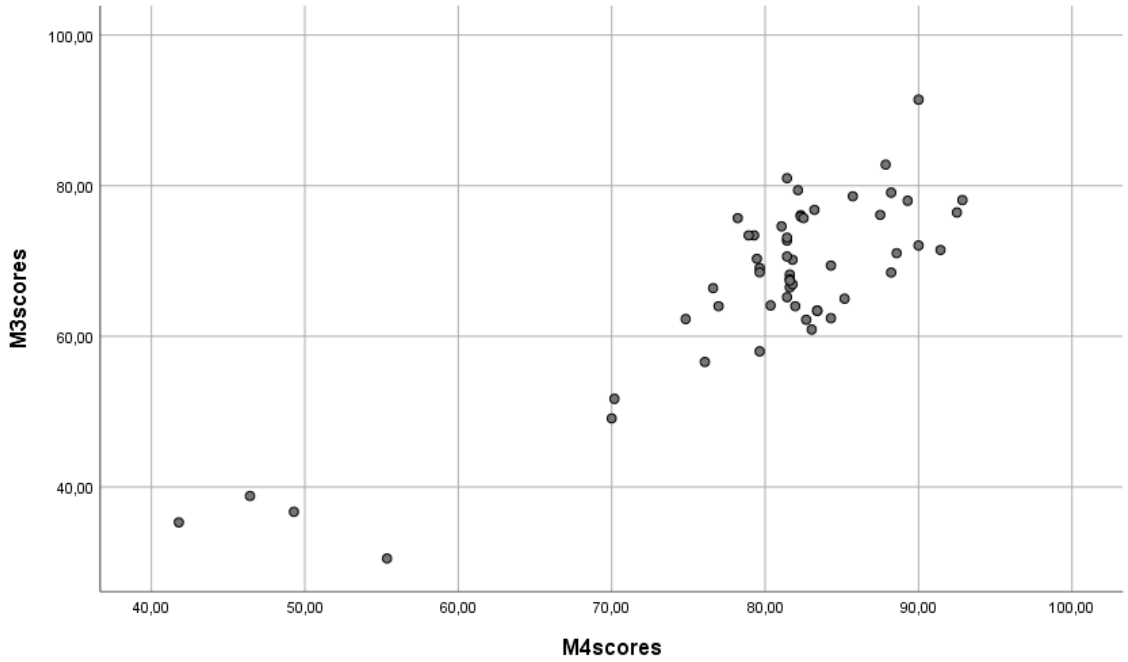
5. Scatterplot incorporating method 2 (three-response dichotomous scoring) and method 3 (aggregate scores)



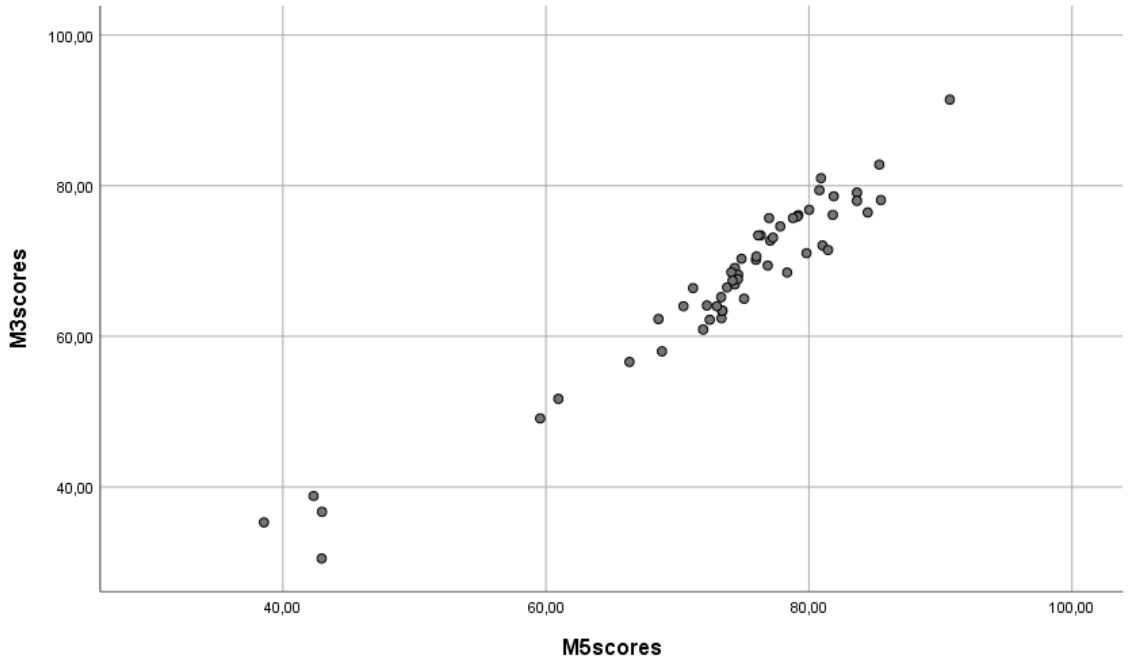
6. Scatterplot incorporating method 2 (three-response dichotomous scoring) and method 4 (distance from experts' modal choices)



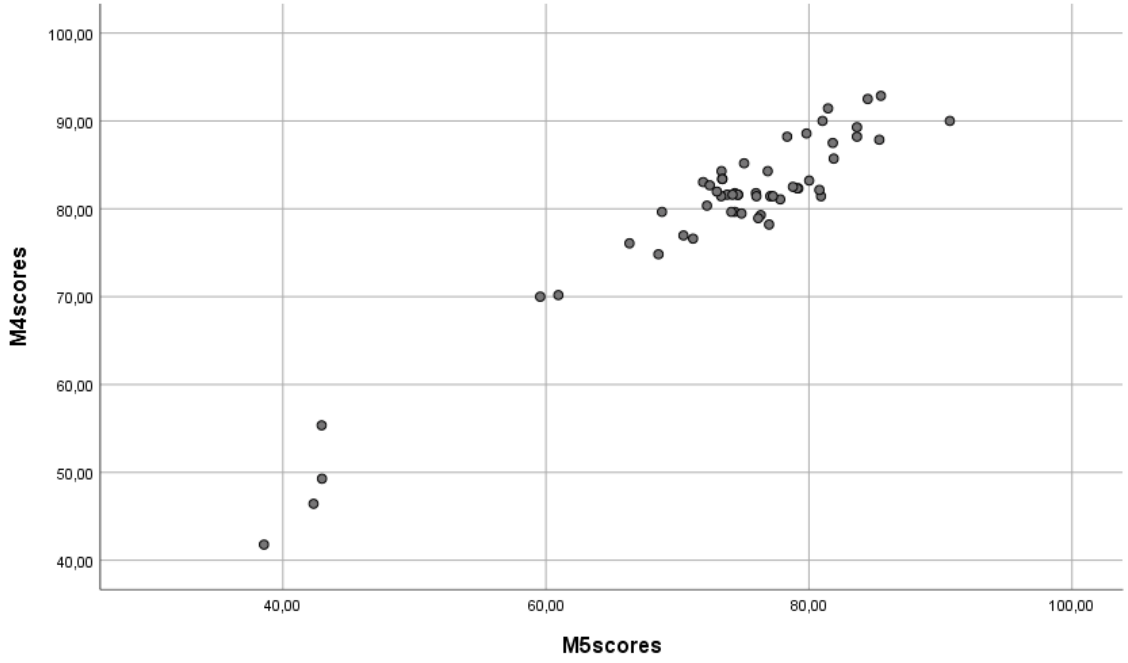
7. Scatterplot incorporating method 2 (three-response dichotomous scoring) and method 5 (combination of aggregate scores and distance from experts' modal choices)



8. Scatterplot incorporating method 3 (aggregated scores) and method 4 (distance from experts' modal choices).



9. Scatterplot incorporating method 3 (aggregated scores) and method 5 (combination of aggregate scores and distance from experts' modal choices).



10. Scatterplot incorporating method 4 (distance from experts' modal choices) and method 5 (combination of aggregate scores and distance from experts' modal choices)